International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 3, May-Jun 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Architectures for Accelerated AI: A Survey of Platforms from Data Centers to Vision-Based Systems

Ananya R, Anshu Naikodi, Archana C K, D Sathya Preetham

Department of Electronics and Communication Engineering,
Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India

ABSTRACT

The research evaluates modern developments in artificial intelligence (AI) and machine learning (ML) about their use in hardware acceleration platforms and data center systems and live systems including autonomous vehicle technology and recommendation engines. Efficient architectural designs for AI chips need emphasis because the industry is expected to generate $70 billion revenue by 2026. The paper shows how artificial intelligence training platforms at Facebook Zion scale up to edge computing designs such as RNNAccel. The research reviews both YOLOv3 and SSD-ResNet with CenterNet within object detection models while investigating the low-memory solution HarDNet. The paper demonstrates how machine learning has merged with vision-based autonomous systems through analyses of navigation integration. This demonstrates the developing combination between hardware systems and optimization software. Multiple research documents show that AI technologies rapidly expand through various computational settings.

How to cite this paper: Ananya R | Anshu Naikodi | Archana C K | D Sathya Preetham "Architectures for Accelerated AI: A Survey of Platforms from Data Centers to Vision-Based Systems" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-3, June 2025, pp.675-680, URL: www.ijtsrd.com/papers/ijtsrd80044.pdf

IJTSRD80044

INTRODUCTION

The author investigates how Artificial Intelligence and Machine Learning techniques integrate within VLSI Very Large Scale Integration design for each design step. The paper starts with an extensive evaluation of VLSI design stages plus an introduction to basic AIML principles when used for VLSI implementation. The paper demonstrates how AIML functions throughout the design process to achieve better results. The paper discusses the major barriers to AIML integration in VLSI design together with the problems linked to insufficient data access and system scalability. The paper demonstrates how AIML technology has the ability to fully transform contemporary integrated circuits through improved performance and operational efficiency. The paper demonstrates that AI through its human-inspired cognitive framework already solved various complex issues across various domains.[1] Many central products and services at Facebook utilize machine learning as its foundation. The following report demonstrates how machine learning operates globally through its hardware and software platforms. Facebook runs numerous applications which demand a wide spectrum of machine learning models because their duties differ substantially. The system stack encounters different impacts due to varying levels of system performance. Facebook experiences data distribution difficulties to its training systems because many stored data points contain machine learning elements. The workloads bring both GPU requirements for training combined with CPU needs for extensive real-time performance. Solving current and new challenges in this field requires continuous joint work among developers from fields which include algorithms and software programming and hardware architecture.[2]

Cloud platforms which employ GPUs and FPGAs already provide successful acceleration of data-intensive operations and result in improved power efficiency together with performance benefits. This paper investigates the essence of ASIC Clouds which represent specialized datacenters developed using

enormous arrays of ASIC (Application-Specific Integrated Circuit) accelerators. Since their initial resistance due to higher initial costs and less flexibility ASIC Clouds have achieved successful deployment through their implementation by commercial Bitcoin network operators at large scale. This document begins by examining Bitcoin mining ASIC Clouds because they function as the largest entities in this category. The design phase of the paper develops three additional ASIC Cloud systems for YouTube video transcoding alongside Linux mining and CNN inference operation. The system performance evaluation demonstrates 2–3 orders of magnitude superior TCO than CPU- or GPU-based power models. Our work produces a methodology which produces optimal ASIC Cloud Servers through the given accelerator design. Data extraction from place-and-routed circuits and computational fluid dynamics simulations leads to a comprehensive brute-force search which optimizes the ASIC and DRAM subsystem and motherboard and power delivery system and cooling system and operating voltage and chassis design simultaneously. Datacenter-level parameters serve as vital factors for choosing the TCO-efficient design from the Pareto-optimal set of options. The paper examines both economic viability factors and NRE impacts on total project feasibility when building an ASIC Cloud infrastructure.[3]. Li-passivation in zigzag GaN nanoribbons significantly modifies their electronic properties, enhancing Fermi velocity and reducing effective mass to improve carrier mobility. DFT investigations further show strong gas adsorption and charge transfer, highlighting their potential as high-performance nanosensors[4-5]. Contemporary Convolutional Neural Networks use floating-point linear algebra for implementation. Lower-precision Neural Networks (NNs) are gaining popularity because they require much less memory as well as fewer computational resources which is crucial for powerlimited applications. The decreased precision results in a minor decrease of accuracy levels. We analyze the accuracy-throughput relationship which occurs when using different NN architectural designs alongside various numerical precision levels.The proposed training method enables dimensional reduction through precision quantization to obtain accurate results during reduced precision inference. The research demonstrates a numerical evaluation of how data precision affects hardware efficiency results. Our study generated multiple important discovery points from the experimental data. The MNIST dataset achieved 99% accuracy best with 32-bit floating-point parameters that demonstrated better hardware efficiency than using 1-bit precision during the same test. When applying fixed-point arithmetic representation with 2-bit and 4-bit precision it becomes possible to achieve superior accuracy rates and processing speed on datasets that range from MNIST to CIFAR-10. The most efficient balance emerges from using 4-bit precision to execute AlexNet while processing the ImageNet dataset.[6]

The paper explains how VLSI employs AI and machine learning (AIML) technology for different stages of the design process. The paper explains VLSI design stages before discussing AIML fundamentals for VLSI system development. The paper moves on to discuss AIML implementation across the entire design workflow. The manuscript explores the limitations and issues which arise from implementing AIML for VLSI design while identifying data availability and scalability problems. The study demonstrates how AIML-enabled methodologies can boost performance together with operational efficiency of contemporary integrated circuit systems. Artificial intelligence (AI) which models human intelligence serves as a solution for numerous problems in various fields.[7]. The paper urges adoption of “scale-in” instead of “scale-out” by creating specialized hardware solutions which position numerous custom processors and memory systems and interconnects on single or dual boards per rack unit. The proposed architecture promises to boost recommender system performance by delivering between 12 and 62 times higher inference throughput as well as training throughput of 12 to 45 times better when compared to the DGX-2 AI platform. The paper assesses performance through a DLRM Facebook case study by examining the impact of primary hardware design elements which consist of memory architecture, communication latency and bandwidth together with interconnect topology. The research investigates AI accelerator and computing platform weaknesses by engaging with demanding usage situations.[8]. The Deep learning recommendation models (DLRMs) stand as Meta's most vital AI workload because they need the maximum computational power throughout their data centers. This document presents the hardware-software design of Neo which serves as a platform for distributed high-performance training operations of large-scale Deep Learning Recommendation Models. Neo utilizes 4D parallelism technology that combines embedding operations with table-wise and row-wise as well as column-wise processing along with data parallelism at scale. The system contains hybrid kernel fusion and software-managed caching and quality-preserving compression technologies which optimize embedding execution time and memory usage. The custom hardware platform ZionEX exists

as a co-developed system with Neo to enhance communication patterns essential for executing extensive DLRM training operations. With the deployment of Neo across 128 GPUs scattered among 16 ZionEX nodes the solution manages to deliver 40× superior performance than current alternatives for training 12-trillion-parameter DLRM models from production pipelines.[9]

This paper investigates the role of machine learning (AIML) and artificial intelligence in VLSI design while showing their applications throughout the entire design process. The paper starts by illustrating the VLSI design workflow structure before introducing AIML concepts as they advance in this application domain. The central section of the work demonstrates how AIML operates throughout the VLSI design workflow to maximize performance effectiveness. The paper evaluates the obstacles and data scale issues which arise when employing AIML for VLSI applications. The paper demonstrates how AIML technology creates revolutionary possibilities for the enhancement of contemporary integrated circuit design processes. The paper shows artificial intelligence derived from human intellectual processes that now applies successfully through numerous domains to handle complicated challenges.[10] .

DFT-based studies demonstrate that Indium Nitride nanoribbons can effectively detect gases like CO, CO_2, NO, and NO_2 due to notable charge transfer and band structure modulation. Similarly, Scandium Nitride monolayers show strong adsorption sensitivity toward toxic gases such as NH_3, AsH_3, BF_3, and BCl_3. Zigzag silicon carbide nanoribbons exhibit enhanced gas sensing performance through improved electronic response to hazardous gas molecules, making them promising for advanced sensor applications[11-13].

The research analyzes AIML applications development in relation to VLSI design along with their impact throughout each design stage. The research first introduces the VLSI design process with its stages before moving to AIML explanations for this domain. The paper investigates how AIML techniques function at multiple phases of VLSI design and demonstrates their benefits for achieving enhanced performance alongside improved efficiency. The work evaluates crucial problems and data restrictions that emerge during application. The paper demonstrates how AIML brings transformative benefits to present-day integrated circuit design. The paper establishes artificial intelligence demonstrates effectiveness because its foundation in human intelligence interpretation enables solutions for challenging problems across multiple domains.[14].

The latest network structures ResNet along with MobileNet and DenseNet prove their accuracy abilities at lower multiply-accumulate operations (MACs) and smaller model sizes. The metrics used for assessment do not necessarily provide accurate indications about actual inference timing. Memory traffic related to feature map access costs turns out to play a substantial role in increasing inference latency in high-resolution applications focusing on object detection and semantic segmentation. Fundamental to the development of efficient networks we designed the Harmonic Densely Connected Network which optimizes MACs and memory traffic for maximum efficiency. Our model decreases inference time by 35% when we compare it against FC-DenseNet-103 and DenseNet-264 and ResNet-50 and ResNet-152 and SSD-VGG reduces the inference time by 36% along with 30% and 32% and a substantial 45% respectively. Our architecture reduces the memory traffic overhead which leads to close correlation between inference latency and memory traffic because we utilize the Nvidia profiler and ARM Scale-Sim tools for validation. Memory traffic needs to serve as a primary design element for developing neural networks meant for edge-based applications with highresolution requirements.[15].

Keypoint-based detection techniques produce numerous incorrect bounding boxes when there are no adequate analytical methods applied to cropped regions for detailed examination. The proposed method uses visual signals from inside target areas to boost detection precision through efficient processing requirements. The single-stage keypoint-based CornerNet detector has been evolved into CenterNet which depicts objects with three keypoint values instead of two. This adjustment enables better precision levels and recall values. The detection quality enhancement comes from two processing units called cascade corner pooling and center pooling which are specifically designed for this purpose. The two newly introduced modules augment context information retrieved from object corners and central regions together. CenterNet reaches a 47.0% average precision (AP) on MS-COCO that surpasses all current one-stage detectors by at least 4.9%. The detection system of CenterNet operates at higher speeds than leading two-stage detectors and achieves similar accuracy levels.[16]. The YOLO object detection system receives multiple enhancements which lead to significant performance improvements. Both network design updates and the development of an increased yet more precise network architecture enabled major performance improvements. The model size increase does not affect its speed performance. YOLOv3 operates at 22 milliseconds

while detecting objects with 28.2 mAP when processing images at 320×320 resolution which matches SSD accuracy but runs three times faster. YOLOv3 reaches top performance when using the AP50 metric with a 0.5 Intersection over Union threshold by delivering 57.9 AP50 through 51 milliseconds of execution on a Titan X GPU. YOLOv3 delivers 57.5 AP50 accuracy at a speed of 198 milliseconds while RetinaNet operates at 57.5 AP50 accuracy but requires 198 milliseconds thus YOLOv3 is 3.8 times faster with equivalent accuracy[17].

Object detection datasets have become more common while variety increases which has improved detection and recognition technology learning abilities. The process of manual annotation becomes both time-consuming and labor-intensive because target objects typically appear as small objects within uniform background environments. The effectiveness of traditional neural networks gets restricted by these detected obstacles. This paper presents a modified network architecture design which deepens the system to improve detection of dangerous goods within different background settings. We build our system on SSD while substituting its VGG16 backbone with the more robust ResNet101 network. The application of ResNet-based models proves effective for detecting various dangerous objects using limited datasets whereas maintaining superior performance compared to alternative neural net architectures.[18].

Self-driving vehicles have recently become prominent because they provide users with travel convenience alongside minimal human operator involvement. The main performance criteria for autonomous driving technology consist of signal identification and lane-keeping ability together with environmental situational awareness targeting pedestrians. These capabilities need object detection for their successful operation. The combination of machine learning technology with computer vision systems leads to the development of an automated detection system for vehicles and both lanes and pedestrians. The system employs ACF (Aggregate Channel Features) algorithm to detect vehicles and simultaneously calculates their detected distance. The detection system shows successful performance in recognizing lanes and vehicles and pedestrians while achieving better results when compared to current detection methods[19].

Density Functional Theory (DFT) investigations reveal that Cu and Fe doping in boron nitride nanoribbons (BNNRs) significantly enhances their electrical conductivity, making them suitable candidates for nanoscale interconnects in advanced integrated circuits. Ab-initio studies on aluminum nitride nanoribbons (AlNNRs) demonstrate their potential in implementing reconfigurable logic gates due to tunable electronic properties under external stimuli. Additionally, the design of a FinFET-based operational amplifier (Op-Amp) using 22 nm high-k dielectric technology shows promising results in reducing leakage currents and enhancing performance, offering a robust solution for low-power, high-efficiency analog circuit applications[20-22].

High-performance machine learning models need vast training procedures to guarantee optimal results. DLRMs make up more than half of all training operations in Facebook's data centers and therefore we focus on these deep learning recommendation models in this paper. The requirement for training recommendation models goes beyond compute power by needing large instances of memory storage and network-wide bandwidth at both memory endpoints. Complexities in efficient scaling of training operations increase when these models continue to expand in size. Our team created Zion as Facebook's next-generation training platform because Facebook needed to address large-memory needs through CPU-accelerator integration. We provide essential design elements for large-scale distributed training system development in the future[23].

Conclusion

Sections of every industry now experience revolutionary changes due to the gathering forces of artificial intelligence with hardware system designs which span from data center operations to edge computing and autonomous systems. AI workloads of today require advances in both powerful computation along with new approaches for memory systems and data management and system expansion capabilities. The combination of Zion platform with HarDNet and CenterNet algorithms alongside decreased computational precision and hardware-software collaboration enables effective trade-offs between system speed and power utilization and result precision. Persons who successfully put AI into advanced applications including self-driving cars and recommendation engines confirm that hardware-oriented AI system development methods are essential.

References:

[1] Tigadi, A. (2023). Survey On VLSI Design For Artificial Intelligence And Machine Learning Applications.

[2] Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., ... & Wang, X. (2018, February). Applied machine learning at

facebook: A datacenter infrastructure perspective. In *2018 IEEE international symposium on high performance computer architecture (HPCA)* (pp. 620-629). IEEE.

[3] M. Jatkar, K. K. Jha and S. K. Patra, "Fermi Velocity and Effective Mass Variations in ZGaN Ribbons: Influence of Li-Passivation," in IEEE Access, vol. 9, pp. 154857-154863, 2021, doi:10.1109/ACCESS.2021.3128294.

[4] M. Jatkar, K. K. Jha and S. K. Patra, "DFT Investigation on Targeted Gas Molecules Based on Zigzag GaN Nanoribbons for Nano Sensors," in IEEE Journal of the Electron Devices Society, vol. 10, pp. 139-145, 2022, doi:10.1109/JEDS.2022.3144014.

[5] Magaki, I., Khazraee, M., Gutierrez, L. V., & Taylor, M. B. (2016). Asic clouds: Specializing the datacenter. *ACM SIGARCH Computer Architecture News*, *44*(3), 178-190.

[6] Su, J., Fraser, N. J., Gambardella, G., Blott, M., Durelli, G., Thomas, D. B., ... & Cheung, P. Y. (2018). Accuracy to throughput trade-offs for reduced precision neural networks on reconfigurable logic. In *Applied Reconfigurable Computing. Architectures, Tools, and Applications: 14th International Symposium, ARC 2018, Santorini, Greece, May 2-4, 2018, Proceedings 14* (pp. 29-42). Springer International Publishing.

[7] Tate, G. (2019). AI Inference Memory System Tradeoffs. *1st Aug*

[8] *Krishna, S., & Krishna, R. (2020). Accelerating recommender systems via hardware" scale-in". arXiv preprint arXiv:2009.05230.*

[9] *Naumov, M., Kim, J., Mudigere, D., Sridharan, S., Wang, X., Zhao, W., ... & Smelyanskiy, M. (2020). Deep learning training in facebook data centers: Design of scaleup and scale-out systems. arXiv preprint arXiv:2003.09518.*

[10] K. K. Jha, M. Jatkar, P. Athreya, T. M. P. and S. K. Jain, "Detection of Gas Molecules (CO, CO2, NO, and NO2) Using Indium Nitride Nanoribbons for Sensing Device Applications," in IEEE Sensors Journal, vol. 23, no. 19, pp. 22660-22667, 1 Oct.1, 2023, doi:10.1109/JSEN.2023.3307761.

[11] Pratham Gowtham, Mandar Jatkar, DFT based study to sense harmful gases (NH3, AsH3, BF3, BCl3) using Scandium Nitride monolayer for sensing device applications, Micro and Nanostructures, Volume 201, 2025, 208100, ISSN 2773-0123, https://doi.org/10.1016/j.micrna.2025.208100.

[12] Jatkar, M. Improving the sensor capability of zigzag silicon carbide nanoribbon for the detection of harmful gases. Discover Electronics 2, 7 (2025). https://doi.org/10.1007/s44291-025-00047-0.

[13] Tigadi, A. (2023). Survey On VLSI Design For Artificial Intelligence And Machine Learning Applications.

[14] Hung, C. M., & Lin, Y. L. (2011). Three-dimensional integrated circuits implementation of multiple applications emphasizing manufacture reuse. *IET Computers & Digital Techniques*, *5*(3), 179-185.

[15] Chao, P., Kao, C. Y., Ruan, Y. S., Huang, C. H., & Lin, Y. L. (2019). Hardnet: A low memory traffic network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3552-3561).

[16] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6569-6578).

[17] Redmon, Joseph and Farhadi, Ali, "YOLOv3: An Incremental Improvement," arXiv:1804.02767, 2018.

[18] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[19] Mandar Jatkar, P.Y. Mallikarjun,"Optimized Cu/Fe doped Boron Nitride Nanoribbons as nanoscale interconnect: DFT Investigation, Materials Science in Semiconductor Processing, Volume 186, 2025, 109050, ISSN 1369-8001,https://doi.org/10.1016/j.mssp.2024.109050.

[20] Sudhir Rai, Kamal K. Jha, Mandar Jatkar, Ab-initio investigation on aluminum nitride nanoribbons for reconfigurable logic gates, Diamond and Related Materials, Volume 152, 2025, 111966, ISSN 0925-9635, https://doi.org/10.1016/j.diamond.2025.111966.

[21] R. Rambola and M. Jatkar, "An Effective Synchronization of ERP in Textile Industries," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 969-973, doi:10.1109/ICECA.2018.8474686.

[22] Gohilot, N. M., Tigadi, A., & Chougula, B. (2021, May). Detection of pedestrian, lane and traffic signal for vision based car navigation. In 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1-6). IEEE.

[23] Naumov, M., Kim, J., Mudigere, D., Sridharan, S., Wang, X., Zhao, W., ... & Smelyanskiy, M. (2020). Deep learning training in facebook data centers: Design of scaleup and scale-out systems. arXiv preprint arXiv:2003.09518.